Нужно прокачать NER-модель, но как?



Прохор Гладких, gladkikh.p.v@sberbank.ru, 24.11.22

18+
ВЫЖИВШИЕ

18:41
18°
ökko
Фильмы и сериалы
Музыка



OK
SBER BOX





SBER DEVICES

Что сейчас будет?

Автор: Прохор Гладких

Должность: Lead Data Scientist

Опыт: DL, Python, C++, Scala, Java, Objective-C

Опыт: 10 лет

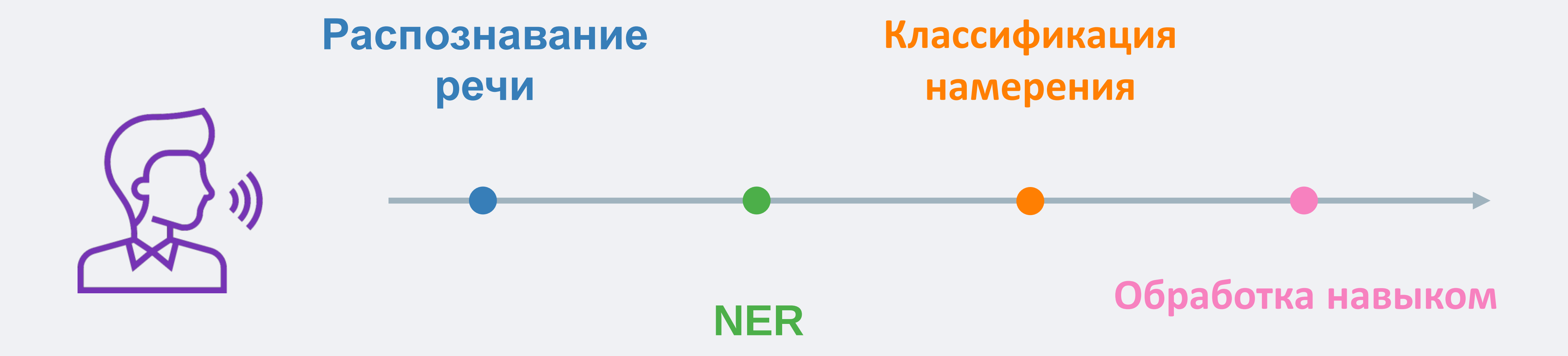
Длительность доклада: 30 мин

Вопросы: 15 мин после выступления



SBER DEVICES

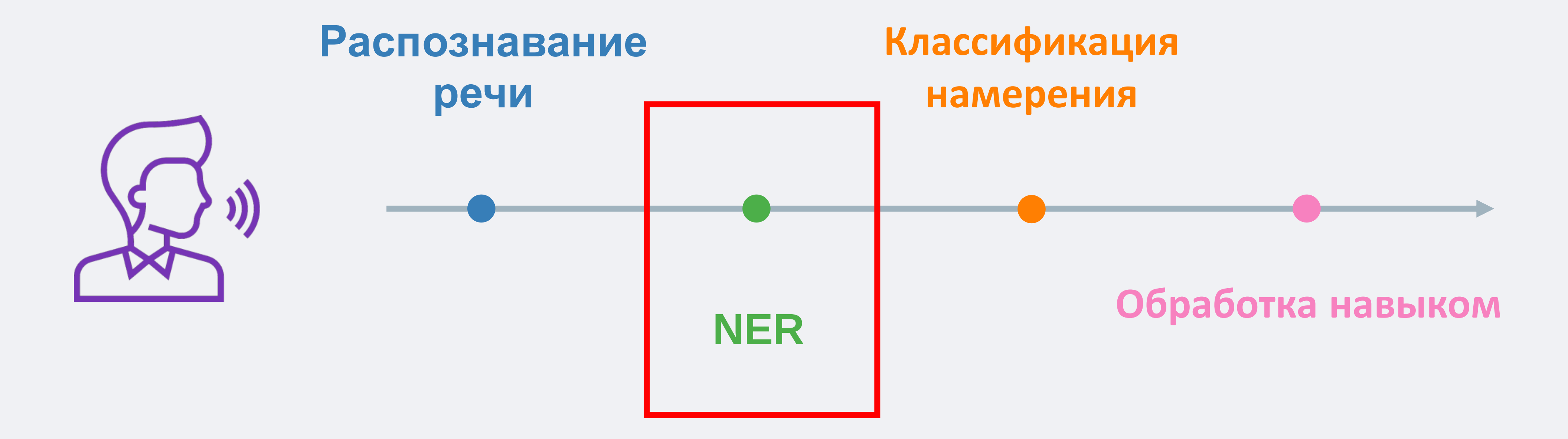
Салют





HL++ HighLoad++ 2022

Салют


Распознавание
речи
Классификация
намерения
NER
Обработка навыком


SBER DEVICES


HL
HighLoad++
2022

Постановка задачи NER*

Сбер – крупнейший банк в России, Центральной и Восточной Европе, а также один из лидирующих международных финансовых институтов.

12 ноября, 1841, Николай I, император Всероссийский, подписал указ о создании сберегательных касс. Эта дата считается днем рождения Сбербанка.

Дата

Геолокация

Организация

Человек

Титул

* - Named Entity Recognition



О чем этот доклад?

1. Введение
2. Улучшение процесса разметки и контроль качества
3. Как избежать распространенных проблем при подготовке датасета?
4. Инструмент анализа ошибок модели
5. На чем подняли качество
6. Вопросы



SBER DEVICES

На чем мы замеряем качество модели?

- test?





На чем мы замеряем качество модели?

- test: **97** f1
- Репутационная корзина: **95** f1
- Частотная корзина: **11** f1
- Частотная корзина GEO: **70** f1

- Еще 7 корзин…

Найди аптеки рядом

Включи звуки для засыпания

Автовокзал



SBER DEVICES

На чем мы замеряем качество модели?

Лучше замерять качество NER на нескольких корзинах

- test: **97** f1
- Репутационная корзина: **95** f1
- Частотная корзина: **11** f1
- Частотная корзина GEO: **70** f1

- Еще 7 корзин…



SBER DEVICES

О чем я НЕ буду рассказывать?

- Как писать train-loop
- Как ставить эксперименты и проводить A/B



Улучшение процесса разметки и контроль качества

LOC — слова или словосочетания, относящиеся к адресам, странам, городам, озерам, рекам, горам, достопримечательностям или адрес целиком

ORG — слова или словосочетания, относящиеся к организациям, компаниям или их категориям

Маршрут до станции метро Охотный ряд



Улучшение процесса разметки и контроль качества

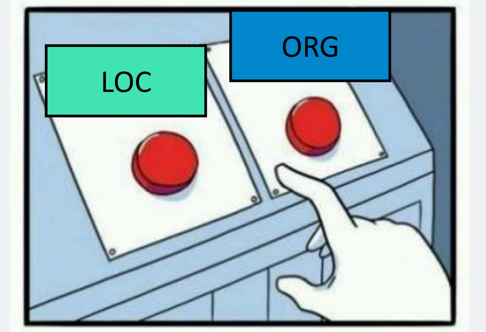
LOC — слова или словосочетания, относящиеся к адресам, странам, городам, озерам, рекам, горам, достопримечательностям или адрес целиком

ORG — слова или словосочетания, относящиеся к организациям, компаниям или их категориям

Маршрут до станции метро Охотный ряд

Маршрут до станции метро Охотный ряд

Маршрут до станции метро Охотный ряд



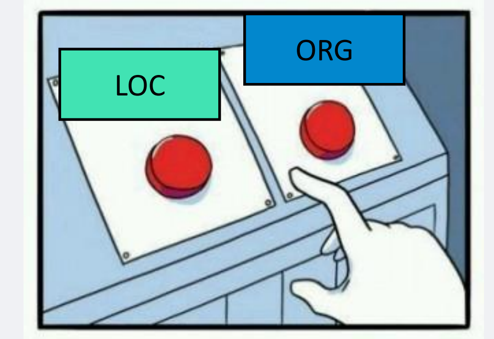




Улучшение процесса разметки и контроль качества

LOC — слова или словосочетания, относящиеся к адресам, странам, городам, озерам, рекам, горам, достопримечательностям или адрес целиком

ORG — слова или словосочетания, относящиеся к организациям, компаниям или их категориям



Маршрут до станции метро охотный ряд

Маршрут до станции метро Охотный ряд

Маршрут до станции метро Охотный ряд

Маршрут до станции метро Охотный ряд





SBER DEVICES

Улучшение процесса разметки и контроль качества

LOC — слова или словосочетания, относящиеся к адресам, странам, городам, озерам, рекам, горам, достопримечательностям или адрес целиком

ORG — слова или словосочетания, относящиеся к организациям, компаниям или их категориям

Маршрут до станции метро охотный ряд

Маршрут до станции метро Охотный ряд

Маршрут до станции метро Охотный ряд

Маршрут до станции метро Охотный ряд



Улучшение процесса разметки и контроль качества

LOC	— слова или словосочетания, относящиеся к адресам, странам, городам, озерам, рекам, горам, достопримечательностям или адрес целиком
ORG	— слова или словосочетания, относящиеся к организациям, компаниям или их категориям

Куда сходить на массаж?



Улучшение процесса разметки и контроль качества

LOC — слова или словосочетания, относящиеся к адресам, странам, городам, озерам, рекам, горам, достопримечательностям или адрес целиком

ORG — слова или словосочетания, относящиеся к организациям, компаниям или их категориям

Куда сходить на массаж?

Куда сходить на массаж?

Куда сходить на массаж?

Куда сходить на массаж?

SBER DEVICES



Улучшение процесса разметки и контроль качества

LOC — слова или словосочетания, относящиеся к адресам, странам, городам, озерам, рекам, горам, достопримечательностям или адрес целиком

ORG — слова или словосочетания, относящиеся к организациям, компаниям или их категориям

Куда сходить на массаж?

Куда сходить на массаж?

Куда сходить на массаж?

Куда сходить на массаж?

SBER DEVICES



Улучшение процесса разметки и контроль качества

LOC — слова или словосочетания, относящиеся к адресам, странам, городам, озерам, рекам, горам, достопримечательностям или адрес целиком

ORG — слова или словосочетания, относящиеся к организациям, компаниям или их категориям

Найди круглосуточный магазин

SBER DEVICES



Улучшение процесса разметки и контроль качества

LOC — слова или словосочетания, относящиеся к адресам, странам, городам, озерам, рекам, горам, достопримечательностям или адрес целиком

ORG — слова или словосочетания, относящиеся к организациям, компаниям или их категориям

Найди круглосуточный магазин

Найди круглосуточный магазин

Найди круглосуточный магазин

SBER DEVICES



Улучшение процесса разметки и контроль качества

LOC — слова или словосочетания, относящиеся к адресам, странам, городам, озерам, рекам, горам, достопримечательностям или адрес целиком

ORG — слова или словосочетания, относящиеся к организациям, компаниям или их категориям

Найди круглосуточный магазин

Найди круглосуточный магазин

Найди круглосуточный магазин



SBER DEVICES

Улучшение процесса разметки и контроль качества

LOC — слова или словосочетания, относящиеся к адресам, странам, городам, озерам, рекам, горам, достопримечательностям или адрес целиком

ORG — слова или словосочетания, относящиеся к организациям, компаниям или их категориям

Путь до продуктового круглосуточного магазина

SBER DEVICES



Улучшение процесса разметки и контроль качества

LOC — слова или словосочетания, относящиеся к адресам, странам, городам, озерам, рекам, горам, достопримечательностям или адрес целиком

ORG — слова или словосочетания, относящиеся к организациям, компаниям или их категориям

Путь до продуктового круглосуточного магазина

Путь до продуктового круглосуточного магазина
Путь до продуктового круглосуточного магазина
Путь до продуктового круглосуточного магазина
Путь до продуктового круглосуточного магазина



Улучшение процесса разметки и контроль качества

LOC — слова или словосочетания, относящиеся к адресам, странам, городам, озерам, рекам, горам, достопримечательностям или адрес целиком

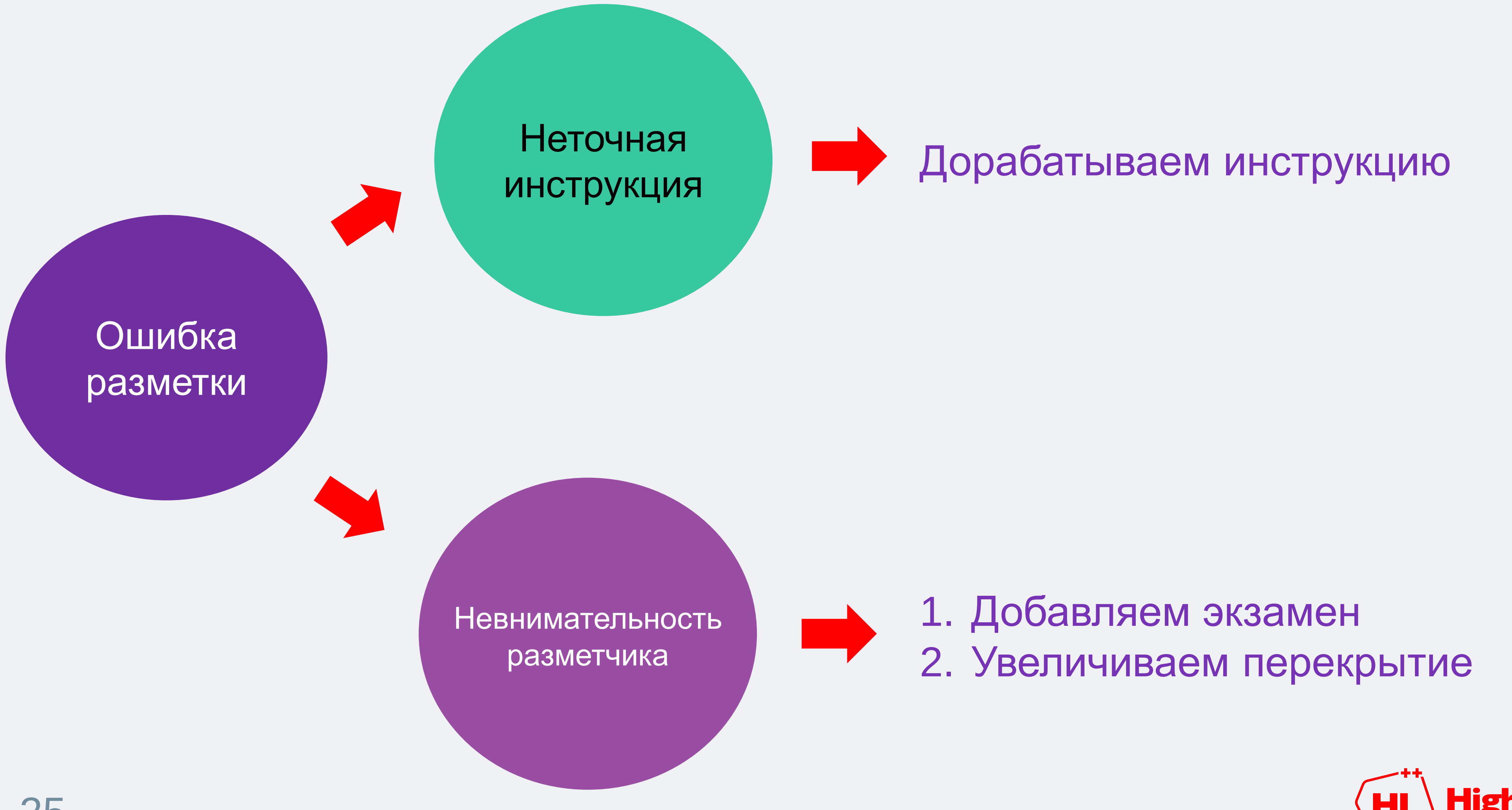
ORG — слова или словосочетания, относящиеся к организациям, компаниям или их категориям

Путь до продуктового круглосуточного магазина

Путь до продуктового круглосуточного магазина
Путь до продуктового круглосуточного магазина
Путь до продуктового круглосуточного магазина
Путь до продуктового круглосуточного магазина



Улучшение процесса разметки и контроль качества



Дорабатываем инструкцию

1. Добавляем экзамен
2. Увеличиваем перекрытие

Перекрытие в разметке







Включи песню Королевство кривых группы Пикник

Перекрытие в разметке







1. Включи песню Королевство кривых группы Пикник
2. Включи песню Королевство кривых группы Пикник
3. Включи песню Королевство кривых группы Пикник
4. Включи песню Королевство кривых группы Пикник
5. Включи песню Королевство кривых группы Пикник



Согласованность

40%

1. **Включи песню Королевство кривых группы Пикник**
2. **Включи песню Королевство кривых группы Пикник**
3. **Включи песню Королевство кривых группы Пикник**
4. **Включи песню Королевство кривых группы Пикник**
5. **Включи песню Королевство кривых группы Пикник**



SBER DEVICES

Улучшение процесса разметки и контроль качества

Доработка инструкции

Разметка

Анализ
согласованности

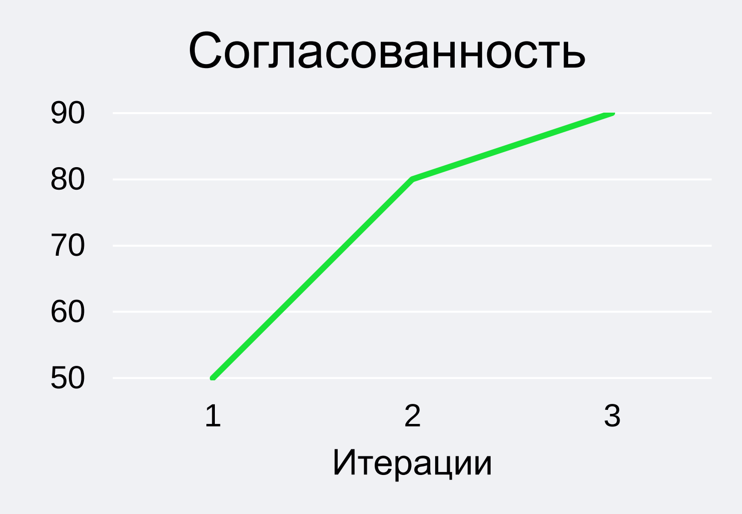


Улучшение процесса разметки и контроль качества

Доработка инструкции

Разметка

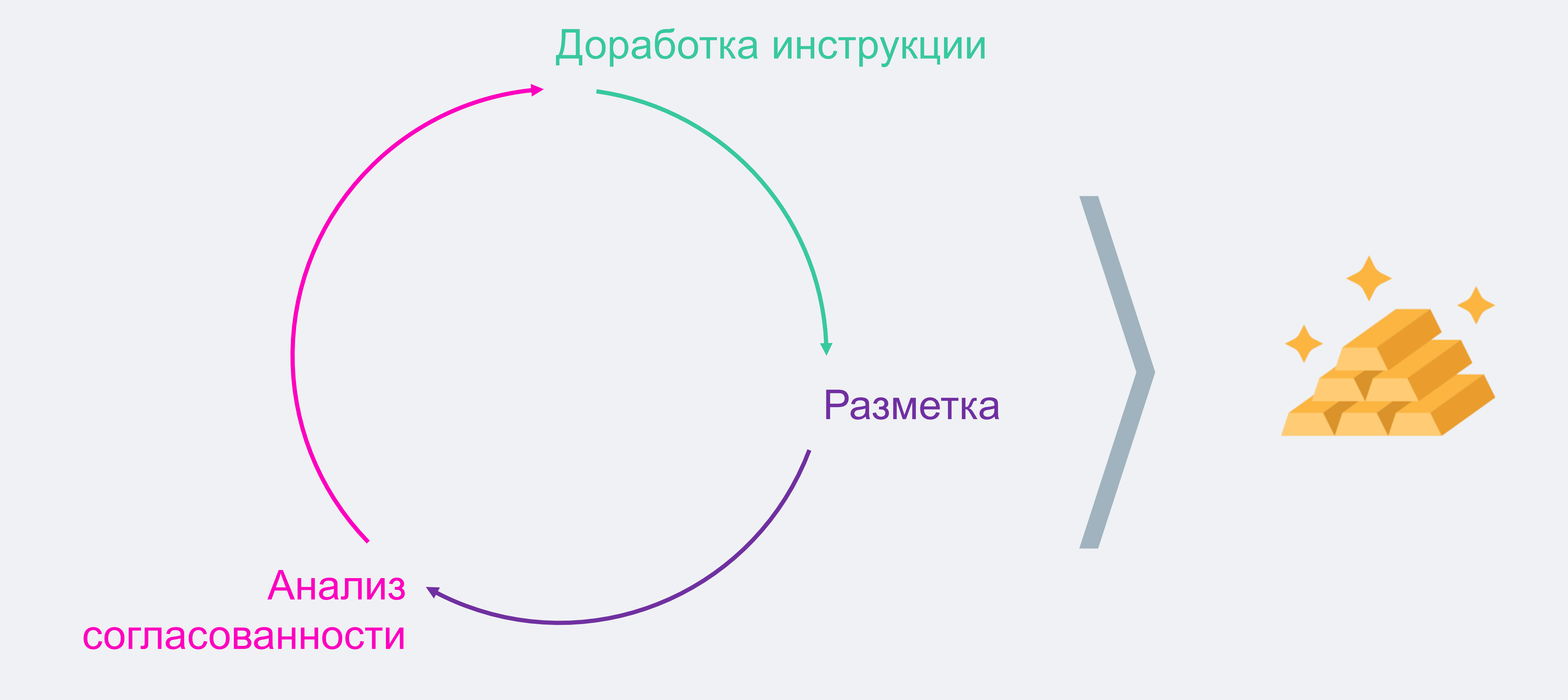
Анализ согласованности





SBER DEVICES

Улучшение процесса разметки и контроль качества



Картинка: Flaticon.com

Как предобработать датасет NER перед обучением?

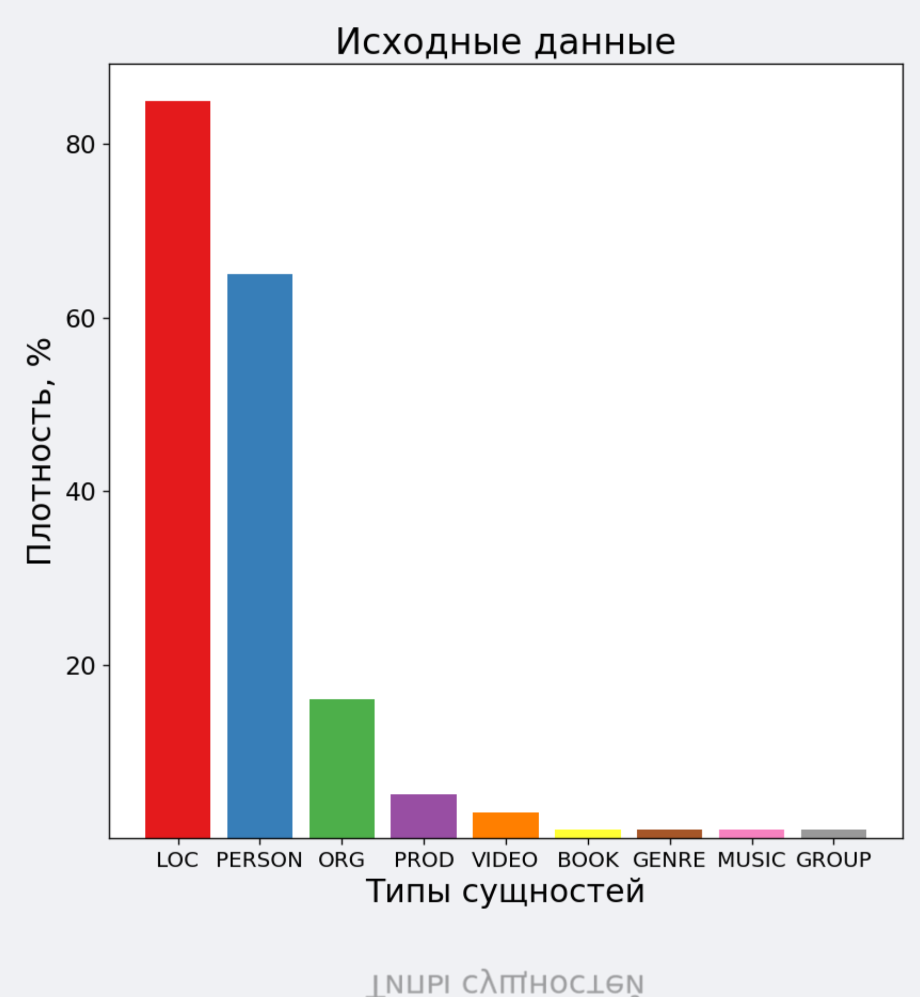
1. Дедупликация текстов

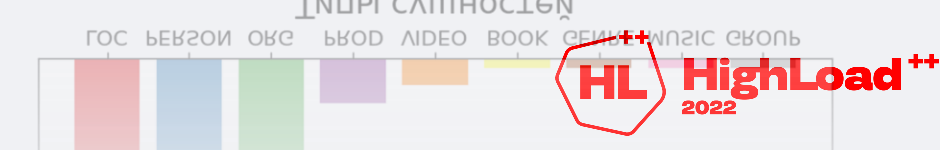


SBER DEVICES

Как предобработать датасет NER перед обучением?

1. Дедупликация текстов





SBER DEVICES

Как предобработать датасет NER перед обучением?

1. Дедупликация текстов
2. Negative Under Sampling


```json
{
  "LOC": 64587,
  "<NO_ENTITY>": 57905,
  "PERSON": 45352,
  "ORG": 15623,
  "PROD": 5537,
  "VIDEO": 2976,
  "BOOK_PERSON": 1974,
  "LOC_ORG": 1511,
  "GENRE": 1101,
  "MUSIC_PERSON": 655,
  "GROUP": 579,
  "MUSIC": 548,
  "GROUP_MUSIC": 491,
  "PERSON_VIDEO": 366,
  "GENRE_PERSON": 257,
  "ORG_PERSON": 243,
  "GENRE_VIDEO": 96,
  "GENRE_LOC": 86,
  "BOOK_GENRE_PERSON": 82,
  "ORG_PROD": 46,
  "GROUP_PERSON": 37,
  "LOC_PERSON": 27,
  "LOC_PROD": 23,
  "ORG_VIDEO": 17,
  "BOOK": 7,
  "GENRE_PERSON_VIDEO": 5,
  "LOC_VIDEO": 4,
  "LOC_ORG_PERSON": 3,
  "GROUP_MUSIC_PERSON": 3,
  "LOC_ORG_PROD": 2
}
```


"GROUP": 579,
"MUSIC": 548,
"GROUP_MUSIC": 491,
"PERSON_VIDEO": 366,
"GENRE_PERSON": 257,
"ORG_PERSON": 243,



SBER DEVICES

Как предобработать датасет NER перед обучением?

1. Дедупликация текстов
2. Negative Under Sampling
3. Стратифицированное разбиение на train/valid/test

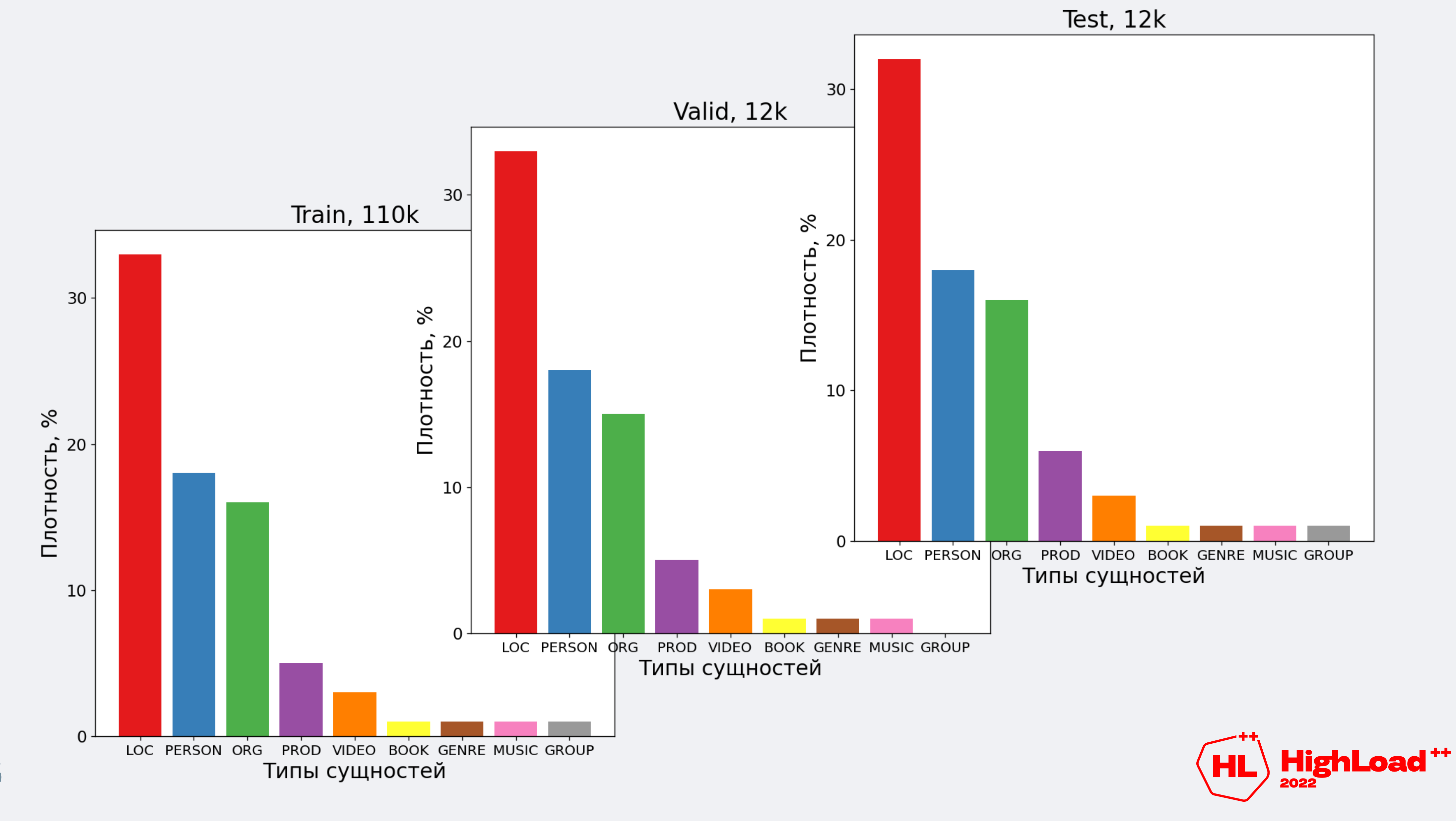
```json
{
  "LOC": 64587,
  "<NO_ENTITY>": 57905,
  "PERSON": 45352,
  "ORG": 15623,
  "PROD": 5537,
  "VIDEO": 2976,
  "BOOK_PERSON": 1974,
  "LOC_ORG": 1511,
  "GENRE": 1101,
  "MUSIC_PERSON": 655,
  "GROUP": 579,
  "MUSIC": 548,
  "GROUP_MUSIC": 491,
  "PERSON_VIDEO": 366,
  "GENRE_PERSON": 257,
  "ORG_PERSON": 243,
  "GENRE_VIDEO": 96,
  "GENRE_LOC": 86,
  "BOOK_GENRE_PERSON": 82,
  "ORG_PROD": 46,
  "GROUP_PERSON": 37,
  "LOC_PERSON": 27,
  "LOC_PROD": 23,
  "ORG_VIDEO": 17,
  "BOOK": 7,
  "GENRE_PERSON_VIDEO": 5,
  "LOC_VIDEO": 4,
  "LOC_ORG_PERSON": 3,
  "GROUP_MUSIC_PERSON": 3,
  "LOC_ORG_PROD": 2
}
```


```
"GROUP": 579,
"MUSIC": 548,
"GROUP_MUSIC": 491,
"PERSON_VIDEO": 366,
"GENRE_PERSON": 257,
"ORG_PERSON": 243,
```


HL HighLoad++ 2022

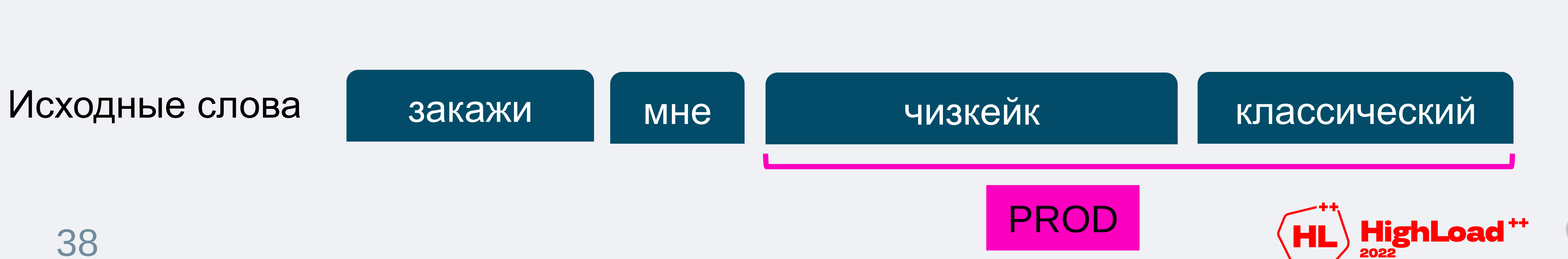


Результат подготовки датасета к обучению




Ваша
NER-модель
ошиблась

Инструмент анализа ошибок модели



Инструмент анализа ошибок модели



Добавить в train

Разработать инструмент анализа ошибок

Исходные слова | закажи | мне | чизкейк | классический

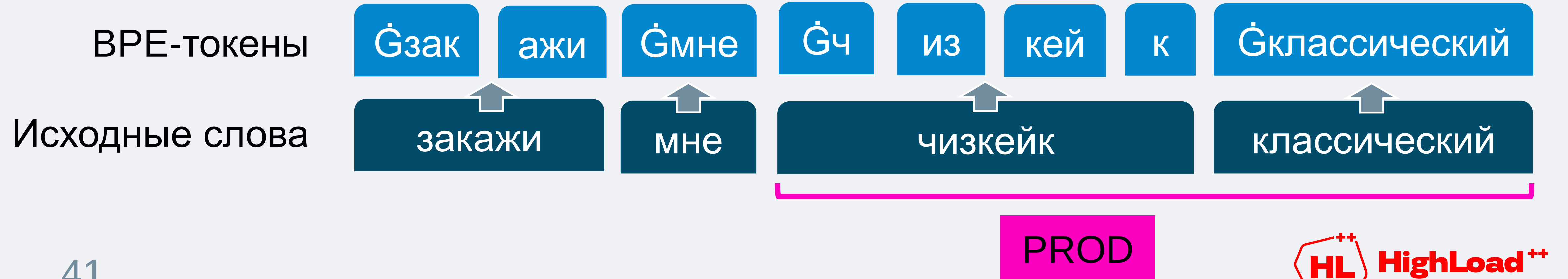
PROD



Инструмент анализа ошибок модели



Инструмент анализа ошибок модели



Инструмент анализа ошибок модели

Инструмент анализа ошибок модели



Инструмент анализа ошибок модели

$R_{\text{сущности}}$

Faiss Index

Pooling

Векторы кодировщика

$R_{\text{Ġзак}}$ $R_{\text{ажи}}$ $R_{\text{Ġмне}}$ $R_{\text{Ġч}}$ $R_{\text{из}}$ $R_{\text{кей}}$ $R_{\text{к}}$ $R_{\text{Ġклассический}}$

Roberta

Embeddings

$E_{\text{Ġзак}}$ $E_{\text{ажи}}$ $E_{\text{Ġмне}}$ $E_{\text{Ġч}}$ $E_{\text{из}}$ $E_{\text{кей}}$ $E_{\text{к}}$ $E_{\text{Ġклассический}}$

BPE-токены

Ġзак ажи Ġмне Ġч из кей к Ġклассический

Исходные слова

закажи мне чизкейк классический

PROD

Инструмент анализа ошибок модели



```
train_error_searcher.run(bucket, k_search=20)
```



Инструмент анализа ошибок модели. Пример работы

Правильно:

Найди круглосуточные банкоматы

Предсказание модели:

Найди круглосуточные банкоматы

Train:

- Найди банкоматы
- Построить маршрут до ближайшего банкомата
- Проложи путь до ближайшего банкомата
- Покажи маршрут до ближайшего банкомата
- Подскажите адрес ближайшего банкомата
- Добрый вечер! Скажите, в Абхазии есть банкоматы Сбербанка?



Инструмент анализа ошибок модели. Пример работы

Правильно:

Найди круглосуточные банкоматы

Предсказание модели:

Найди круглосуточные банкоматы

Train:

- Найди банкоматы
- Построить маршрут до ближайшего банкомата
- Проложи путь до ближайшего банкомата
- Покажи маршрут до ближайшего банкомата
- Подскажите адрес ближайшего банкомата
- Добрый вечер! Скажите, в Абхазии есть банкоматы Сбербанка?

Ошибки разметки



SBER DEVICES

Инструмент анализа ошибок модели. Пример работы

Правильно:

Поехали до кремля

Предсказание модели:

Поехали до кремля

Train:

- Построй дорогу до кремля
- Построить маршрут до кремля
- Кратчайшая дорога до кремля
- Как мне добраться до кремля?
- Помоги мне проложить маршрут до кремля
- На каком виде транспорта можно доехать до кремля?



SBER DEVICES

Инструмент анализа ошибок модели. Пример работы

Правильно:

Поехали до кремля

Предсказание модели:

Поехали до кремля

Train:

- Построй дорогу до кремля
- Построить маршрут до кремля
- Кратчайшая дорога до кремля
- Как мне добраться до кремля?
- Помоги мне проложить маршрут до кремля
- На каком виде транспорта можно доехать до кремля?

Ошибки разметки

SBER DEVICES



Что думает модель NER про это слово?

Придуманный запрос:

Построй-ка маршрут мне до Луны



Что думает модель NER про это слово?

Придуманный запрос:

Построй-ка маршрут мне до Луны



```
train_index.search("Построй-ка маршрут мне до <Луны>")
```



Что думает модель NER про это слово?

Придуманный запрос:

Построй-ка маршрут мне до луны



Ближайшие запросы:

- Маршрут до галактики
- Построй маршрут до горизонта
- Построить маршрут до континента
- Проложи маршрут до планеты
- Как пройти до континента?
- Сделай дорогу до континента
- Построить маршрут до планеты
- Построй мне маршрут до ленты пешком
- Как доехать до ленты маршрут
- Как доехать до глобуса?



Что думает модель NER про это слово?

Придуманный запрос:

Закажи доставку гуся в яблоках



Что думает модель NER про это слово?

Придуманный запрос:

Закажи доставку гуся в яблоках



```
train_index.search("Закажи доставку <гуся в яблоках>")
```




Что думает модель NER про это слово?

Придуманный запрос:

Закажи доставку гуся в яблоках

PRODUCT

Ближайшие запросы:

- Купить колбасу клинскую мини салями сырокопченую
- Найди джей севен фрукты целиком томат стопроцентный сок 0.97 литра
- Найди мороженое пломбир вологодский
- Хочу заказать батончик picnic грецкий орех
- Найди круглозерный рис камолино теплые традиции
- Найди колбасу солями по-черкизовски
- Купить йогурт коломенское молоко малина груша
- Заказать кофе coffesso classico



Как удалось прокачать качество?

NER f1
90.8 (+20.3)

- Расчет метрик на разных корзинах с ПРОМ-а
- Исправлена проблема лика, дубликатов и дисбаланса классов в датасете
- Переработана инструкция по разметке (100500 итераций)
- Размечен полностью новый датасет
- Разработан инструмент анализа ошибок NER
- Новая модель на основе Roberta-large
- Другие манипуляции с датасетом





SBER DEVICES

Как удалось прокачать качество?

NER f1
90.8 (+20.3)

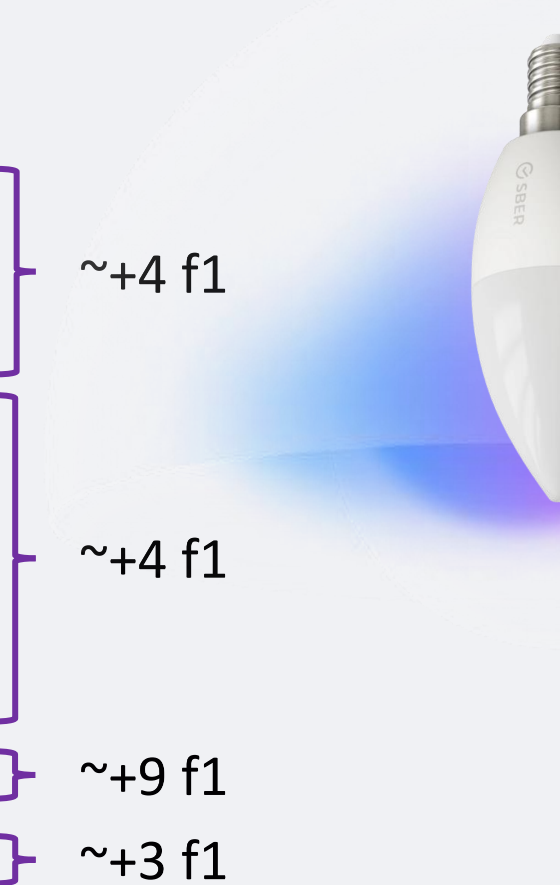
- Расчет метрик на разных корзинах с ПРОМ-а
- Исправлена проблема лика, дубликатов и дисбаланса классов в датасете

~+4 f1

- Переработана инструкция по разметке (100500 итераций)
- Размечен полностью новый датасет
- Разработан инструмент анализа ошибок NER

~+4 f1

- Новая модель на основе Roberta-large — ~+9 f1
- Другие манипуляции с датасетом — ~+3 f1





SBER DEVICES

Salute AI Day

30 ноября

Приходите пообщаться с нашими командами SberDevices

Подробности в чате встречи



Время вопросов

Обратная связь



Прохор Гладких

 prohor33

 prohor-gladkikh

 prohorgladkikh

 prohorgladkikh

Ссылка на слайды:

shorturl.at/emEGW

